

Empirical Analysis of the Online Rating Systems

Xin-Yi Lu,¹ Jian-Hong Lin,¹ Qiang Guo,¹ and Jian-Guo Liu^{1, a)}

Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, P. R. China

(Dated: 29 October 2015)

This paper is to analyze the properties of evolving bipartite networks from four aspects, the growth of networks, the degree distribution, the popularity of objects and the diversity of user behaviours, leading a deep understanding on the empirical data. By empirical studies of data from the online bookstore Amazon and a question and answer site Stack Overflow, which are both rating bipartite networks, we could reveal the rules for the evolution of bipartite networks. These rules have significant meanings in practice for maintaining the operation of real systems and preparing for their future development. We find that the degree distribution of users follows a power law with an exponential cutoff. Also, according to the evolution of popularity for objects, we find that the large-degree objects tend to receive more new ratings than expected depending on their current degrees while the small-degree objects receive less ratings in terms of their degrees. Moreover, the user behaviours show such a trend that the larger degree the users have, the stronger purposes are with their behaviours except the initial periods when users choose a diversity of products to learn about what they want. Finally, we conclude with a discussion on how the bipartite network evolves, which provides guideline for meeting challenges brought by the growth of network.

PACS numbers: 64.60.aq, 82.56.Lz, 07.05.Tp

I. INTRODUCTION

In past years, evolving networks have arose the interests of lots of researchers, who study the evolving networks in diverse fields,^{1,2} such as person-to-person communication,^{3,4} one-to many information dissemination,^{5,6} neural and brain networks,^{7,8} and ecological networks.^{9,10} Derek Price first studied the growth of citation networks for scientific papers.¹¹ He found that the more citations a paper received, the more chances it would be cited in the future, which was called "cumulative advantage" by him.¹¹ Barabasi and Albert later studied the World Wide Web, where they found a power-law degree distribution^{12,13} and proposed the well-known model of network growth, named as Preferential attachment (PA). A great number of other researchers such as Strogatz¹⁴, Huang¹⁵, Liu¹⁶, Daz¹⁷ also do empirical studies to learn about dynamics of complex networks, building solid foundation for analyzing the development of bipartite networks.

Interaction between a large number of entities are common in natural and man-made systems. In order to study these interaction, real systems are mathematically represented as graphs, consisting of pairs of nodes and a set of edges, where nodes represent the entities and the edges represent the interaction between entities.¹⁸ However, the previous studies mainly focus on networks with nodes of same type and pay little attention to the evolving networks with other structure, like bipartite networks. Bipartite networks conclude nodes in two types: $U = \{u_1, u_2, \dots, u_n\}$ and $V = \{v_1, v_2, \dots, v_m\}$ where edges only exist between nodes of different types,¹⁸ as shown

in Figure 1. The most common evolving bipartite networks are e-business networks, which update frequently and combine with the human dynamics. Thus, in this paper, we study the evolving rating networks deeply from several aspects, such as the growth of networks, the user degree distribution, the popularity of objects and the diversity of user behaviours, by examining the empirical data of Amazon and Stack Overflow. Medo once used

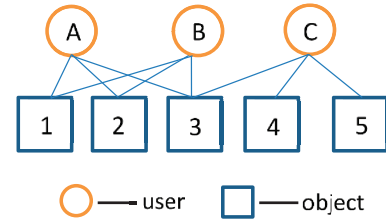


FIG. 1. (Color online) A bipartite network with nodes in two types: $U = \{A, B, C\}$ and $V = \{1, 2, 3, 4, 5\}$. Each edge represents the rating that the object acquired from the user.

a model based on Preference Attachment to analyze the citation network.¹⁹ According to his study, we apply the model in investigating the evolution of popularity for objects and use the information entropy to study the user behaviours. Studying properties of such networks can help us better understand the evolving direction of bipartite networks and capture the features of user behaviours.

The rest of this paper is organized as follows. In Sec. II, we analyze statistical properties of data sets of Amazon book ratings and the Stack Overflow favourite posts to display the basic situation of these systems. In Sec. III, we propose a network growth model to study the evolving popularity of objects. Then we present the features of user behaviours by calculating the information

^{a)}liujg004@ustc.edu.cn

entropy in Sec. IV. Finally, we discuss our results and the direction of future study in Sec.V.

II. STATISTICAL PROPERTIES OF AMAZON BOOK RATINGS AND STACK OVERFLOW FAVOURITE POSTS

We analyze two data sets, Amazon and Stack Overflow, to study the statistical properties of evolving bipartite networks, such as the growth of networks and the degree distribution of users. The Amazon is an on-line shopping service which encourages its users to rate the commodities they buy. The data set of Amazon we analysed is the Amazon book ratings, including 2005409 ratings delivered by 99622 users on 645055 books, from 31st.May,1996 to 15th.Sep,2005. The Stack Overflow is a question and answer website of the Stack Exchange Network, which invites its users to mark their favourite posts. The data set of Stack Overflow has 1301942 ratings, given by 545196 users on 96680 posts, from 2002 to 2005. Table I gives the basic statistical properties of Amazon and Stack Overflow. Every entry of these two data sets records the user ID and the object ID with time stamp. We count up the number of ratings in Amazon for each day from March 2001 to April 2001 to study temporal dynamics in the data. The result shows a weekly pattern for the number of ratings of Amazon, shown as Figure 2, which presents the dynamical pattern of human activities.

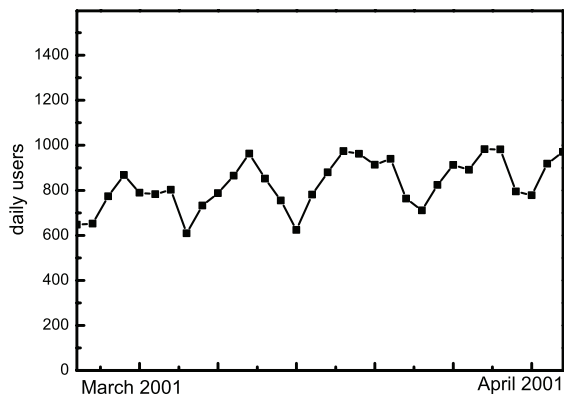


FIG. 2. (Color online) Daily number of users of Amazon from March 2001 to April 2001. Similar pattern can be observed in the whole network of Amazon.

A. The growth of networks

The Amazon and Stack Overflow are studied as evolving bipartite networks with two types of nodes, users U and objects V . The edges between users U and objects V present the ratings the users giving to the objects.²⁰ We focus on whether the rating events exist regardless the value of ratings. The large size data with time stamp

for long time offer us the chance to reveal the growth of networks, from which we can get an overall view of the development for Amazon and Stack Overflow.

Figure 4 shows the total number of ratings every day in the data sets of Amazon and Stack Overflow respectively, which grow exponentially with time.

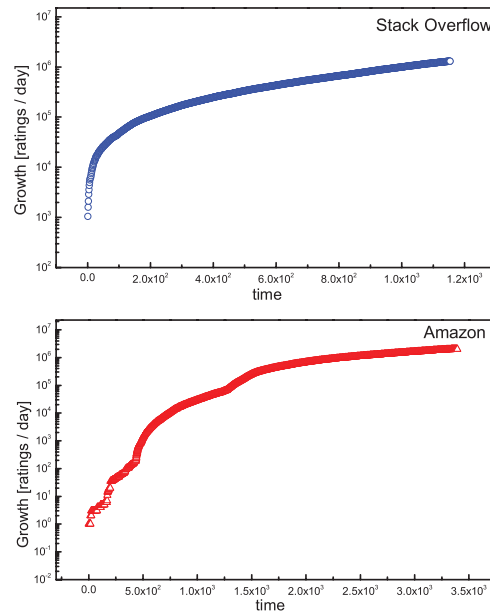


FIG. 3. (Color online) The total number of ratings in the data sets of Amazon and Stack Overflow respectively, which increases exponentially with time.

B. Degree distribution

The degree distribution of users for Amazon and Stack Overflow follow a power law with an exponential cutoff. The degree of the user in the bipartite network is the total number of ratings that the user gives to objects and the degree distribution is the probability distribution of user degrees over the whole network.²¹ To study the aggregated structure, we examine the cumulative degree distributions of users, the result of which fits a power law with exponential cutoff by the method modified by Clauset *et al*,²²

$$F(x) \sim x^{-\alpha} e^{-\lambda x} \quad (1)$$

Figure 4 shows the cumulative degree distributions of users for the evolving networks Amazon and Stack Overflow. Table II shows the value of parameters α and λ according to the Eq.(1).

Systems in maturity period are systems which have stable interactions within themselves or with their environments. From the result we can find that, for systems in maturity period, most users are small-degree users while few have large degrees. Users seek gain when rating or finding favourite posts with limited time or energy. In economics, this kind of gain is called the utility

TABLE I. Basic statistical properties of Amazon and Stack Overflow, including the number of users u , the number of the objects v , time range and time span.

Network	users	objects	Rating records	Time range	Time span(day)
Amazon	99622	645055	2005409	31st.May,1996-15th.Sep,2005	3394
Stack Overflow	545195	96680	1301942	2002-2005	1154

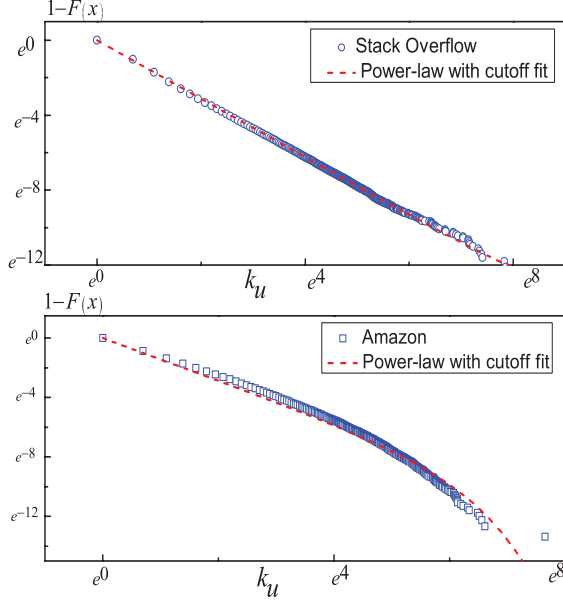


FIG. 4. (Color online) The cumulative degree distributions of users for the evolving networks Amazon and Stack Overflow. For Amazon, the dashed curves fitting well to the cumulative degree distributions of users are with parameters $\alpha = 1.4274$, $\lambda = 0.00327$. For Stack Overflow, the dashed fitting curves are with parameters $\alpha = 1.5593$, $\lambda = -0.00112$.

TABLE II. Fitting parameters of user degree distribution for Amazon and Stack Overflow.

Network	mean α	variance α	λ	variance λ
Amazon	1.4274	0.0077	0.0033	0.00153
Stack Overflow	1.5593	0.0012	-0.00112	0.00012

and the marginal utility is the gain from an increase, or the loss from a decrease.²³ When the utility of users achieves equality (marginal utility equals the marginal cost),²⁴ users gradually cease the activities of ratings or marking the favourite posts due to the law of diminishing marginal utility.²⁵

III. THE EVOLVING POPULARITY OF OBJECTS

The popularity of objects is decided by complex reasons and evolves with the network growth. Evolving net-

works change with time, not only by adding or removing links, but also by adding or removing nodes, which are closest to natural networks. The first evolving network model was BA model proposed by Barabási Albert to study the scale free networks,²⁰ including two significant concepts, growth and preferential attachment. Based on the BA model, we use a model to study the evolving popularity of objects. Regarding to the system science, systems have varied states with different structures. In this paper, we focus on systems in maturity period where the evolution of popularity for objects can be more representative.

A. The model for popularity of objects

In this section, we propose a dynamical model to investigate the evolving popularity of objects. According to the well-known model of network growth, Preference Attachment, the probability of a node v_i with degree k_i acquiring new contacts is

$$P(v_i) = k_i / \sum_{j=1}^N k_j \quad (2)$$

In our model, we use Relevance R_i to show a regular way how popularity of objects evolve, where the degree of object i at time t is defined as $k_i(t)$. Relevance R_i is the ratio between the real number of ratings one object received and the expected number of ratings the object will receive. We assume that $X(t, \Delta t)$ is the number of new ratings added to objects in the network during the next Δt days. What need to be noticed is that the total number of objects varied with time. So m is a variable with time t , used to indicate the number of existed objects. The expected number of ratings that object i will receive during next Δt can be denoted as

$$\Delta k_i(t, \Delta t) = X(t, \Delta t) k_i(t) / \sum_{j=1}^m k_j(t) \quad (3)$$

And we use $\Delta K_i(t, \Delta t)$ to represent the real number of ratings that object i received. Thus, the Relevance R_i is shown as Eq.(4).

$$R_i(t, \Delta t) = \Delta K_i(t, \Delta t) / \frac{X(t, \Delta t) k_i(t)}{\sum_{j=1}^m k_j(t)} \quad (4)$$

However, we ignore the initial time of networks when each object receive its first rating. The development of networks at initial time is complex which is related to the "cold boot". To better clearly display how the popularity of objects evolves, we assume that the $k_i(t) \geq 1$ and no time periods have no ratings, that is, $C(t, \Delta t) \neq 0$.

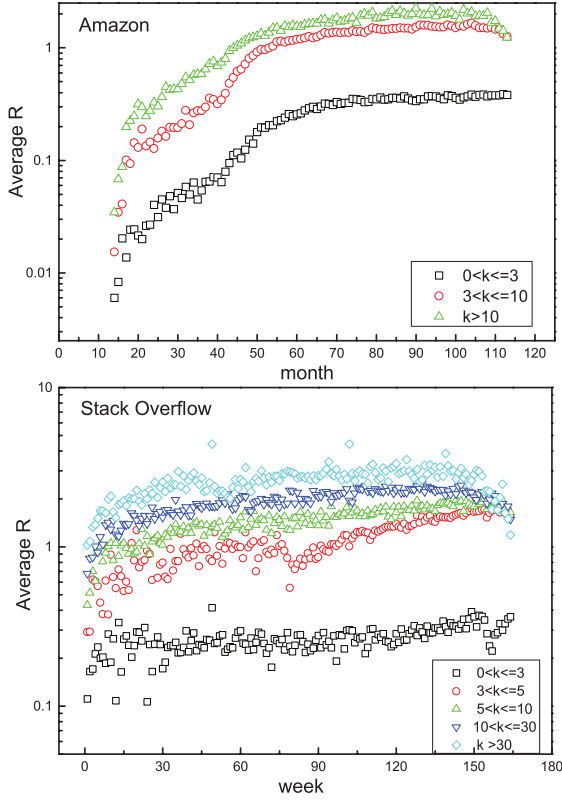


FIG. 5. (Color online) The evolution of objects' popularity for Amazon and Stack Overflow. For Amazon, the objects are divided into three groups by their degrees with $\Delta t = 30\text{days}$. When the system goes into a maturity period, the average Relevance $R(i)$ of the group $k > 10$ fluctuates around 2 while the Relevance R of the group $0 < k \leq 3$ is about 0.36. For Stack Overflow, the objects are divided into five groups based on their degree with $\Delta t = 7\text{days}$. The value of the Relevance R of the group $k > 30$ is about 3 as the group $0 < k \leq 3$ only fluctuates around 0.27.

B. The empirical study of the evolving popularity of objects for Amazon and Stack Overflow

We apply the empirical data of Amazon and Stack Overflow to our model. For Amazon data, we first divide all objects into three groups depended on their degrees as $0 < k \leq 3$, $3 < k \leq 10$, and $k > 10$. Then, we separate the data into 114 time windows to see how it evolves with time. For Stack Overflow data, we group the objects into five types based on their degrees as $0 < k \leq 3$, $3 < k \leq 5$, $5 < k \leq 10$, $10 < k \leq 30$, and $k > 30$ and divide the sequence of time into 165 time windows. We calculate the average Relevance R_i of different groups for Amazon and Stack Overflow, respectively. It can be clearly seen that the average Relevance R_i increases exponentially at the initial time. Then, R_i starts to increase slowly until the networks evolve to a certain stage, the maturity period. It is the time when networks have stable interactions within themselves or with their environments. So R_i will fluctuate around a certain value in that period.

However, the R_i of objects with large degree will decrease after a certain time, since the number of ratings received by objects with large degree arrives a saturate level.

From the above results, we find that not all the evolution of popularity for objects fit the PA method and only objects with not too large or small degree fit the PA method well. The number of ratings given to objects in next period is not completely driven by their current popularity, which is influenced by other exterior reasons, like the environment of the market and limitation of the user's number. The popularity of objects in each group exhibits heterogeneous fitness values which increase with the growth of evolving bipartite networks, shown as Figure 5. The larger the degree of one object is, the more ratings it will receive compared to the expected number of ratings calculated by Eq.(3). Finally, due to the limitation of the user's number, the number of ratings received by large-degree objects would first achieve the saturate level so that the objects' popularity will gradually decay. Figure 5 shows an evolving process of the objects' popularity.

IV. THE EVOLVING DIVERSITY OF USER BEHAVIOURS

In this section, we discuss the features of user behaviours by calculating the information entropy. To operate and manage an open system, one of the key points is to understand user behaviours. We need to predict the user behaviours by collecting and analyzing the potential signals so that we can ensure the systems are under control. Moreover, for e-business systems, learning about the user behaviours can help them to upgrade the experience of the users,^{26–28} make personal recommendation and provide customized advice to users.

A. The information entropy

Entropy is introduced to measure disorder or uncertainty. Shannon entropy was first proposed by Claude Shannon in 1948 to measure the unpredictability of information content.²⁹ Warren Weaver extended the meaning of the "information" mentioned by Shannon. He pointed out that the word information in communication theory is not related to what you do say, but to what you could say. That is, information is a measure of the diversity of one's choice.³⁰ For instance, a choice H can lead to n results while the possibility of each result is expressed as p_1, p_2, \dots, p_n . The information entropy of this choice can be shown as Eq.(5)

$$H = H(p_1, p_2, \dots, p_n) = \sum_{i=1}^n p_i \log_2 \frac{1}{p_i} \quad (5)$$

As we known, the larger the information entropy is, the more uncertain the system is. And the information entropy can be used to measure the diversity of user

behaviours.³¹ We group the objects and users by their degrees that each group of users and objects has same degree value. Then, we measure the diversity of users in each group to display the features of user behaviours for the whole system. The information entropy of users with degree k is shown as Eq.(6). We assume k is the degree of user; $n(k)$ is the total number of objects with different degrees, chosen by users with degree k ; p_i is the probability that objects with degree i chosen in total ratings times.

$$H(k) = \sum_{j=1}^{n(k)} p_i \log_2 \frac{1}{p_i} \quad (6)$$

p_i can be calculated according to the PA model, as shown in Figure 7. That is, the possibility of objects in each group can be calculated depending on each group's degree. K_i is the total degree of objects with degree i and N is the count of the users with different degrees.

$$p_i = K_i / \sum_{j=1}^N K_j \quad (7)$$

Thus, the larger the information entropy is, the more diverse the user behaviours are. It is appropriate to display the features of user behaviours in bipartite networks by the information entropy.

B. The empirical study of user behaviours for Amazon and Stack Overflow

We do empirical study of the user behaviours in the data sets Amazon and Stack Overflow by calculating the information entropy. First, we group the users and objects by their degree for Amazon and Stack Overflow respectively. Then we employ each data sets to Eq.(6). To better display the features of user behaviours, we normalized the information entropy by the principle of maximum entropy. As we known, it is a natural property for systems to develop toward to disorder. So the choice with largest entropy can best represent the state of other choices^{32,33} and it is the time when p_i is same. The $\theta(k)$ is the normalized information entropy, shown as Eq.(8).

$$\theta(k) = \sum_{j=1}^{n(k)} k_i \log_2 \frac{1}{p_i} / \log_2 n(k) \quad (8)$$

Figure 6 shows that $\theta(k)$ abruptly increases as the user degree increases initially, which can be up to nearly 0.92 and 0.98 for Amazon and Stack Overflow respectively, when 1 is the time that no preference exists in user behaviours at all. Then the information entropy decreases gradually as the degree of users increases. The result illustrates a phenomenon that users tend to choose different types of objects when they first enter the network. Then gradually their behaviours are acted with more and more strong purpose, which means that users start to know clearly about what they want. We fit the results by the least square method. Thus, there is an apparent trend that shows the decreasing inclination of the normalized information entropy, that is, the interests of user

behaviours tend to be centralized as their degrees raise. Generally speaking, the interests of large-degree users are more centralized than small-degree users.

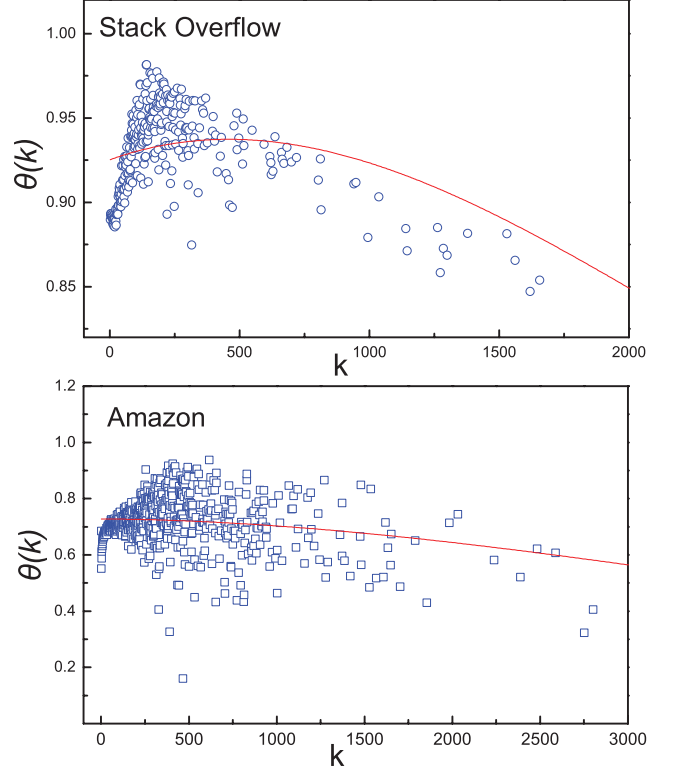


FIG. 6. (Color online) The normalized information entropy $\theta(k)$ of the user with different degree k for Stack Overflow and Amazon.

V. CONCLUSION AND DISCUSSIONS

A. Summary

In this paper, we investigate the evolving properties of systems by empirical study of the Amazon and Stack Overflow. To better display the different aspects of the systems, we abstract the Amazon and Stack Overflow as bipartite networks, studying the development of these systems as well as the weekly pattern of rating behaviours and the degree distribution. We also find that the user degree distributions of Amazon and Stack Overflow follow a power law distribution with an exponential cutoff.

Moreover, we presented a network growth model, in which we can learn about the evolving popularity of objects. Based on the idea of Preference Attachment, the average relevance $R(i)$ is applied to reveal the evolvement of rating process. By empirical analysis, we find that the popularity of objects in each group exhibits heterogeneous fitness values which increases with the growth of evolving bipartite networks. The objects with large degree will receive more ratings compared to the expected

number of ratings based on PA method and vice versa. The objects with not too large or small degree can best fit the PA method. Current popularity cannot be completely accounted as the reason deciding the number of ratings that objects receive. Furthermore, we study the properties of evolving networks from the perspective of users. The information entropy is applied to the data sets of Amazon and Stack Overflow. The users in these two networks present a common phenomenon, that the larger degree the user have, the stronger purpose are with their behaviours. In conclusion, this paper discusses the properties of networks from the aspects of users, objects and rating behaviours, which gives advice to the system operators on how to face the challenges brought by the evolution of bipartite networks.

B. Limitations and future work

Although our paper discusses different aspects of the evolving bipartite networks, it also has a few defects: Firstly, in this work, the network growth model can only display a simple process how the popularity of objects evolve, lacking details analyzing the specific factors that influence the evolvment process. Regardless of the complexity and diversity of the user behaviours, this model cannot be a sufficient basis to explore thoroughly the evolving popularity of objects in bipartite networks. Therefore, the model of network growth needs further research in the future work.

Secondly, from the result of our network growth model, we can see that the relevance $R(i)$ of the groups with largest degree for Amazon and Stack Overflow both decrease after a certain time. As another direction of the future work, we need to explore deeper about when objects with large degree will cease to acquire more ratings. If we can predict the time when the popularity of welcomed objects will decay by monitoring the relative index, it will have remarkable influence on the business research.

Thirdly, it is worth pointing out that we study diversity of user behaviours in the aggregated networks without comparing the difference of the diversity of user behaviours at different time. In this sense, the dynamical user behaviours would be a key issue for our future work.

ACKNOWLEDGMENTS

We thank Kai Yang for useful comments and suggestions.

- ¹P. Holme, J. Saramaki, Temporal networks[J], Physics reports. (519)2012 97-125.
- ²G. Caldarelli, Scale-free networks: complex webs in nature and technology[J], Oxford University Press Catalogue. (2007).
- ³J. P. Eckmann, E. Moses, D. Sergi, Entropy of dialogues creates coherent structures in e-mail traffic[J], Proc. Natl. Acad. Sci. USA. 101(2014) 14333-14337.

- ⁴J. L. Iribarren, E. Moro, Impact of human activity patterns on the dynamics of information diffusion[J], Physical review letters. 103(2009).
- ⁵E. Adar, L. A. Adamic, Tracking information epidemics in blogspace[C], In Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence.(2005) 207-214.
- ⁶D. Liben-Nowell, J. Kleinberg, Tracing information flow on a global scale using Internet chain-letter data, Proc. Natl. Acad. Sci. USA. 105(2008) 4633-4638.
- ⁷O. Sporns, D. R. Chialvo, M. Kaiser, C. C. Hilgetag, Organization: development and function of complex brain networks, Trends in Cognitive Sciences. 8(2004) 418-425.
- ⁸E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nature Reviews Neuroscience. 10(2009).
- ⁹M. Pascual, J. Dunne, Ecological Networks: Linking Structure to Dynamics in FoodWebs. Oxford University Press. Oxford UK. (2006).
- ¹⁰R. V. Sol, J. Bascompte, Self-Organization in Complex Ecosystems[M], Princeton University Press. (2006).
- ¹¹D. J. Price, Networks of Scientific Papers[J], Science. 149(1965) 510-515.
- ¹²A. L. Barabási, R. Albert, Emergence of scaling in random networks[J], Science. 286(1999) 509-512.
- ¹³A. L. Barabási, R. Albert, H. Jeong, Mean-field theory for scale-free random networks[J], Physica A. 272(1999) 173-187.
- ¹⁴S. H. Strogatz, Exploring complex networks[J], Nature. 410(2001) 268-276.
- ¹⁵Z. Huang, D. D. Zeng, H. Chen, Analyzing consumer-product graphs: Empirical findings and applications in recommender systems[J], Management science, 53(2007) 1146-1164.
- ¹⁶J. G. Liu, Y. Z. Dang, Z. T. Wang, Complex network properties of Chinese natural science basic research[J], Physica A. 366(2006) 578-586.
- ¹⁷M. B. Daz, M. A. Porter, J. P. Onnela, Competition for popularity in bipartite networks[J], Chaos. 20(2010).
- ¹⁸J. L. Gross, J. Yellen, eds., Handbook of Graph Theory[M], CRC Press. (2004).
- ¹⁹M. Medo, G. Cimini, S. Gualdi, Temporal effects in the growth of networks[J], Physical review letters. 107(2011).
- ²⁰R. Albert, A. L. Barabási, Statistical mechanics of complex networks[J]. Reviews of modern physics. 74(2002).
- ²¹M. E. J. Newman, The Structure and Function of Complex Networks[J], SIAM review. 45(2003) 167-256.
- ²²A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-law distributions in empirical data[J], SIAM review. 51(2009) 661-703.
- ²³A. Marshall, Principles of Economics, Library of Economics and Liberty[J], Retrieved August. 6(1920).
- ²⁴L. Rittenberg, T. D. Trigarthen, Principles of Microeconomics[M], Flat World Knowledge. (2009).
- ²⁵H. H. Gossen, The laws of human relations and the rules of human action derived therefrom[M], Mit Press. (1983).
- ²⁶T. Zhou, J. Ren, M. Medo, *et al.* Bipartite network projection and personal recommendation[J], Physical Review E. 76(2007).
- ²⁷T. Zhou, Z. Kuscsik, J. G. Liu, *et al.* Solving the apparent diversity-accuracy dilemma of recommender systems[J], Proc. Natl. Acad. Sci. USA. 107(2010) 4511-4515.
- ²⁸J. G. Liu, T. Zhou, Q. Guo, *et al.* Effect of user tastes on personalized recommendation[J], arXiv preprint. (2009).
- ²⁹C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal[J]. 27(1948) 379-423 and 623-656.
- ³⁰C. E. Shannon, W. Weaver, The Mathematical Theory of Communication[M], University of Illinois press. (1949).
- ³¹Y. L. Zhang, J. Ni, Q. Guo, J. G. Liu, Empirical analysis of diversity of online user interests, Application Research Of Computers. 31(2014).
- ³²E. T. Jaynes, Information theory and statistical mechanics[J], Physical review. 106(1957) 620-630.
- ³³E. T. Jaynes, Information theory and statistical mechanics II[J], Physical review. 108(1957) 171-190.